IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 10, No. 2, June 2021, pp. 452~457
ISSN: 2252-8938, DOI: 10.11591/ijai.v10.i2.pp452-457

Performance comparison between naive bayes and k- nearest neighbor algorithm for the classification of Indonesian language articles

Titin Winarti, Henny Indriyawati, Vensy Vydia, Febrian Wahyu Christanto
Faculty of Information Technology and Communication, Semarang University, Indonesia

Article Info

Article history:

Received Mar 22, 2020
Revised Apr 11, 2021
Accepted Apr 22, 2021

Keywords:

Articles classification
Indonesian language articles
K-nearest neighbor
Naive bayes

ABSTRACT

The match between the contents of the article and the article theme is the main factor whether or not an article is accepted. Many people are still confused to determine the theme of the article appropriate to the article they have. For that reason, we need a document classification algorithm that can group the articles automatically and accurately. Many classification algorithms can be used. The algorithm used in this study is naive bayes and the k-nearest neighbor algorithm is used as the baseline. The naive bayes algorithm was chosen because it can produce maximum accuracy with little training data. While the k-nearest neighbor algorithm was chosen because the algorithm is robust against data noise. The performance of the two algorithms will be compared, so it can be seen which algorithm is better in classifying documents. The comes about obtained show that the naive bayes algorithm has way better execution with an accuracy rate of 88%, while the k-nearest neighbor algorithm has a fairly low accuracy rate of 60%.

Corresponding Author:

Titin Winarti
Department of Information Technology and Communication
Universitas Semarang
Soekarno Hatta Street, Semarang, Central Java 50196, Indonesia
Email: titin@usm.ac.id

1. INTRODUCTION

One form of publication is the article. To be accepted in publication, the article must have a content match to the theme of the article where the article will be published. Articles that are related to the theme of the specified article will be reviewed first before the article is accepted or not. While the article containing contents that have too far related to the theme will immediately be rejected. The suitability of the article content and the article theme is the main factor whether or not an article is accepted, [1]. For that reason, we need a document classification algorithm that can classify articles automatically. Each article must have a unique word that can represent the content of the article. This is used as a reference to conduct classification accurately, so there is no need to read the entire article to determine whether the article is accepted or not accepted [2].

Many algorithms can be used to classify articles, including support vector machine (SVM), k-nearest neighbor algorithm, and naive bayes [3]. Research that utilizes the classification algorithm is Irfa, A. A., Adiwijaya, A., and Mubarok, M. S. The research using the k-nearest neighbor method in classifying news article documents [4]. The research of Irfa, resulting in a value of 69.9%. Research on the classification of documents using the k-nearest neighbor method was also conducted by Suharno. In the research conducted by Suharno explained that the comparison of document classification using k-nearest neighbors with feature

selection gets lower accuracy than Irfa's research [5]. Another study was conducted by Mugiardi, G., and Nurwidyantoro, A. The topic was naive bayes algorithm is the most excellent performing classification calculation when the number of features used is more than or equal to 3000 [6]. Other research was conducted by Razi, AR. Classification of Indonesian news articles using the convolutional neural network. This research yields the test results of the system that is built to show that a combination of neural network and word2vec methods provides better results better than the naive bayes method, the value of precision of 96.70% and precision, recall, function reach 96.60%. This study uses 5000 data [7]. The above research study was conducted with a large dataset, around 3000 data.

Research using the naive bayes algorithm has been conducted by M. Ridwan to evaluate performance with large data [8]. From this research, it is known that the naive bayes algorithm successfully performed the classification accuracy. This calculation as were requires a expansive sum of preparing information to decide the assessed parameters required within the classification prepare [9]. Subsequent research was conducted by Winarti *et al*. for analysis on Indonesian texts with large data [10]. The results of this study indicate that the k-nearest neighbor algorithm can provide good performance because the algorithm is resilient to data noise [11]. While the SVM Algorithm has a good level of accuracy but has a long processing time compared to the k-nearest neighbor algorithm [12].

This article will discuss the naive bayes algorithm and k-nearest neighbor to classify Indonesian language articles on small or small amounts of data. The algorithm with the best performance can be used to build an article classification system that can make it easier to choose a theme that matches the articles it has. This system provides an opportunity for an article to be accepted so that it can get publication [13], [14].

2. RESEARCH METHOD

Data and information collection techniques used in this research is by conducting a literature study that will produce secondary data. The data used for classification is abstract of Indonesian language articles. There will be 40 documents in the Indonesian language that will be used. The documents were downloaded from the website. For more details of the workflow of the document, classification shown in Figure 1.

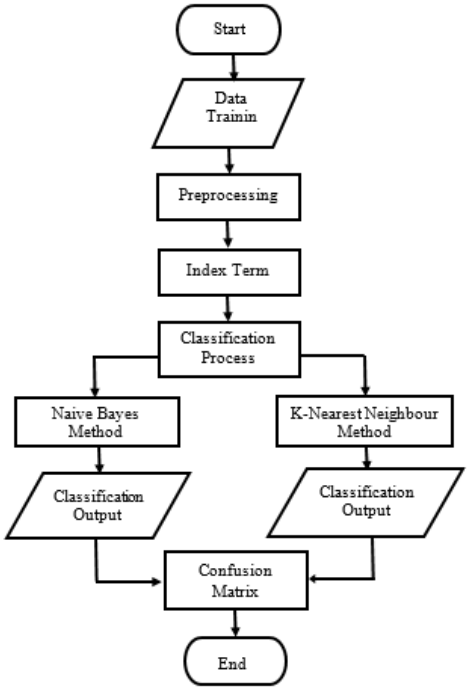

Figure 1. Research stages

2.1. Data training

The classification of articles made there is two stages. The first stage is the learning process or training of a set of articles (training set) and the next stage is the process of classifying articles whose categories are unknown (testing set) based on the knowledge that has been formed from the training set.

2.2. Text preprocessing

Text preprocessing aims to prepare text documents into data ready to be processed in the next process. The stages of preprocessing carried out are Tokenizing stage to decapitate words in documents. Spaces are used as separators among the words. At this stage, filtering will also be done by removing certain characters, such as punctuation. The case collapsing arrange is to alter all capital letters in a record into lowercase letters. The characters other than the a-z letter will be considered as delimiter [15]

2.3. Index term

Index term or word weighting with TF-IDF. Term frequency (TF) and inverse document frequency (IDF) is the most commonly used weighting [16]. TF-IDF algorithm is a way to find the weight of a word (term) in a document [17]. The TF-IDF algorithm combines two ways to calculate its weight, to be specific by calculating the recurrence of event of words in a specific document (TF) and performing an inverse calculation on the document frequency containing the word (IDF) [18]. The inverse document frequency (IDF) calculation is used to calculate the number of terms that function as a measure of the significance level of a term in a document [19].

2.4. Classification process

This article will compare the execution of the naïve bayes and k-nearest neighbor algorithm.

2.4.1. Naive bayes classification (NBC)

Naive bayes may be a basic probabilistic classification algorithm. This algorithm will calculate a set of probabilities by including up the recurrence and combination of values from a given dataset. The naive bayes algorithm considers all attributes in each category don't have a dependency on each other (independent) [20]. The advantage of utilizing naive bayes is that it as it were requires a little sum of preparing information to decide the cruel and fluctuation parameters of the factors required for the classification [21]. Naive bayes is a supervised document classification algorithm which means it requires training data before carrying out the classification process. In the training process, the category (training data) of the documents have been determined, which will then be processed and create information within the shape of likelihood values for each word. This prepare will deliver a word in each report that characterizes the archive in a certain category.

After conducting the training process, the following is the classification process. In this process, the document category (test data) used has not been known yet, so the naive bayes algorithm will search for words in the test data that match with the knowledge in the training data. Then calculate the probability of each document that has been stored in the knowledge at the time of the previous training process, then each category of the documents can be calculated.

2.4.2. K-nearest neighbor

K-nearest neighbor is a supervised algorithm which means it requires training data to classify the object that is the closest in distance [22]. The working standard of k-nearest neighbor is to locate the nearest separation between the information to be assessed with k neighbors in the training data [23]. In the training process, the documents are grouped manually according to predetermined categories. After that, the document will go through the preprocessing stage which will produce a weight for each word in all the training documents. At that point figure the likeness of the test report vector with every one of the preparation archives that have been ordered. The cosine likeness algorithm is utilized to decide the comparability of records of documents [24]. The distance calculation uses a cosine similarity algorithm.

The next step is sorting the distance based on the smallest (the closest) value to the largest (the farthest). Then determine the number of neighbors (k values) that want to be used as a reference for the classification process [25]. From this k value, it can be determined the document category based on the nearest Euclidean distance value each category of the documents can be calculated.

2.5. Confusion matrix

Evaluation of the classification results of Indonesian language articles is carried out using a confusion matrix [26]. This Algorithm represents the results of the classification using a matrix.

3. RESULTS AND ANALYSIS

3.1. Dataset

Dataset used are as many as 40 articles that have been published in the last 2 years, these articles include; information technology, economics, law, and civil engineering articles. Each article is taken as many as 10 articles.

3.2. Preprocessing results

Preprocessing is done so that the information can be prepared for the classification stage. Some of the tokenization and calculation of TF-IDF results in each document category shown in Table 1.

Table 1. Tokenisation and TF- IDF results

	TF-IDF			
Term	Information Technology	Economics	Law	Civil Engineering.
anggaran	0	0	1,795	0
asosiatif	0	0	0	2,556
dividen	0	0	0	1,596
efikasi	0	2,556	0	0
equity	0	0	0	1,596

3.3. Classification results

The testing technique using k-fold cross-validation is very suitable for processing small amounts of data. The working principle is to divide data as much as k subsets, where k is the value of the fold. Then, each of the sub-set will be used as test data from the classification results generated from the k-1 other sub-sets. Each data will be test data 1 time and become training data k-1 times. This study uses the fold value=10 so that the 40 data will be divided into 10 blocks with the same number of training, namely 4 instances. Each data will become testing data 1 time and become training data 3 times (k-1). Some documents of the classification results are shown in Table 2 for the naive Bayes algorithm and in Table 3 for the k-nearest neighbor algorithm.

Table 2. Naive bayes classification

Doc	The real classification	The result classification	Comment
39	Civil Engineering	Civil Engineering	True
2	Information Technology	Information Technology	True
28	Law	Information Technology	False
13	Information Technology	Information Technology	True
34	Economics	Economics	True
4	Information Technology	Information Technology	True
22	Economics	Information Technology	False
15	Economics	Information Technology	False
37	Civil Engineering	Civil Engineering	True
7	Information Technology	Civil Engineering	False

Table 3. K-nearest neighbor classification

Doc	The real classification	The result classification	Comment
39	Civil Engineering	Information Technology	False
2	Information Technology	Civil Engineering	False
28	Law	Information Technology	False
13	Information Technology	Information Technology	True
34	Economics	Economics	True
4	Information Technology	Information Technology	True
22	Economics	Information Technology	False
15	Economics	Economics	True
37	Civil Engineering	Civil Engineering	True
7	Information Technology	Civil Engineering	False

The algorithm will be performed with the k-fold cross validation algorithm which will produce a confusion matrix which can be seen in Tables 4 and 5. Data will have true value if the documents category of the classification results using the naive bayes or k-nearest neighbor algorithm is the same as the actual data category and will have a false value if the class of the classification results is not the same as the actual class. The confusion matrix of each algorithm shown in Tables 4 and 5.

Table 4 shows that a is the article in the information technology category, b is the article in the economics category, c is the article in the law category, and d is the article in the civil engineering category. The articles from 10 documents by the naive Bayes algorithm succeeded in classifying 8 documents according to their class, while the other 2 were incorrectly classified as economics and civil engineering articles. Then, there are articles in the economics category of 10 documents, 2 documents were successfully classified according to their class, while 8 other documents were incorrectly classified as information technology, law, and civil engineering articles. Furthermore, there are articles in the law category, out of 10 documents there are 8 documents classified according to their class correctly, while 2 other documents are incorrectly classified as civil engineering articles. Then there are civil engineering articles as many as 10 documents, and all of these articles are classified according to their class. The test results using the confusion matrix of the k-Nearest Neighbor algorithm shown in Table 5. In this study, the authors set the parameter k=5.

Table 4. Confusion matrix of naive bayes

a	b	c	d	Classified as
8	1	0	1	a= Information Technology
5	2	1	2	b= Economics
0	0	8	2	c= Law
0	0	0	10	d= Civil Engineering

Table 5. Confusion matrix of k-nearest neighbor

a	b	c	d	Classified as
3	7	0	0	a= Information Technology
3	6	0	1	b= Economics
0	8	2	0	c= Law
0	5	0	5	d= Civil Engineering

The confusion matrix of the k-nearest neighbor algorithm shows Table 5, out of 10 documents in the information technology journal category, there are 3 documents classified according to their class, while 7 other documents are incorrectly classified as economics journals. For the 10 documents of the economics category articles, there are 6 documents classified according to their class, while 3 other documents are incorrectly classified as the information technology journal and 1 other document is incorrectly classified as the civil engineering journal. Then the document in the category of law journals, there are 2 documents out of 10 documents classified according to their class, while 8 other documents are incorrectly classified as the economics journal. for the civil engineering journal category, there are 5 documents by their class and 5 other documents are incorrectly classified as economics.

3.4. Test result

Based on 40 documents that have been tested, the results of the calculation of precision, recall, accuracy, and error of each algorithm are obtained. The results of each algorithm testing shown in Table 6. Based on Table 6, it can be seen that the performance of the naive bayes algorithm is better than the k-nearest neighbor algorithm. However, the classification accuracy cannot achieve perfect results in the absence of errors. This is influenced by the amount of test data and training data used and the preprocessing stages carried out. For the naive bayes algorithm, the obtained accuracy is quite good, this is because of the advantages of the naive Bayes algorithm itself, which is capable of doing classification even though it has little training data for estimating its parameters. Whereas the k-nearest neighbor algorithm produces low accuracy, this is because the algorithm is not effective if there is only a small amount of training data.

Table 6. Comparison of performance

Algorithm	Accuracy	Recall	Precision	Error
Naive Bayes	88%	88%	88,9%	12%
k-Nearest Neighbor	60%	60%	64,1 %	40%

4. CONCLUSION

After applying the naive bayes and k-nearest neighbor algorithm to classify articles in the Indonesian language, it is known that the performance of the naive bayes algorithm is superior to the k-nearest neighbor algorithm. It is proven that from the 40 test data used, the naive bayes algorithm was able to classify 28 Indonesian-language articles. While the k-nearest neighbor algorithm can only classify Indonesian-language articles of 16 documents out of 40 test data. This can be influenced by the amount of data used and the stages of preprocessing carried out. Therefore, it is recommended to add data sets and to complete the preprocessing stages such as doing word stemming in further research. The results obtained show that the naive bayes algorithm has better performance with an accuracy rate of 88%, while the k-nearest neighbor algorithm has a fairly low accuracy rate of 60%. So the naive bayes algorithm produces high performance for small and large scale data.

ACKNOWLEDGEMENTS

Thank you to the Semarang University, the Faculty of Information and Communication Technology, and the Institute of Research and Community Service for funding the sustainability of this articles.

REFERENCES

[1] Pambudi, R. Agung, and M. S. Mubarok, "Multi-Label Classification of Indonesian News Topics Using Pseudo Nearest Neighbor Rule," *Journal of Physics: Conference Series*, vol. 1192, no. 1, 2019, Art. no. 012031, doi: 10.1088/1742-6596/1192/1/012031.

[2] W. Gao, S. Oh and P. Viswanath, "Demystifying Fixed K -Nearest Neighbor Information Estimators," *IEEE Transactions On Information Theory*, vol. 64, no. 8, pp. 5629-5661, 2018, doi: 10.1109/Tit.2018.2807481.

[3] P. T. Noi, and M. Kappas, "Comparison of Random Forest, K-Nearest Neighbor, And Support Vector Machine Classifiers For Land Cover Classification Using Sentinel-2 Imagery," *Sensors*, vol. 18, no. 1, p. 18, 2018, doi: 10.3390/S18010018.
[4] A. A. Irfa, Adiwijaya, and M. S. Mubarok, "Classification Of Indonesian News Topics Using K-Nearest Neighbor," *Eproceedings of Engineering*, vol. 5, no. 2, pp. 3631-3640, 2018.
[5] C. F. Suharno, M. A. Fauzi, & R. S. Perdana, "Indonesian Text Classification On Online Complaint Documents Using The K-Nearest Neighbors And Chi-Square Methods," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 1, no. 10, pp. 1000-1007, 2017.
[6] G. Mugiardi, & A. Nurwidyantoro, "Comparison Of Feature Selection Methods On Classification Of Indonesian News Articles Using Naive Bayes," Doctoral Dissertation, Universitas Gadjah Mada, 2017.
[7] R. A. Razi, "Indonesian Language News Article Classification Using Convolutional Neural Network," Doctoral Dissertation, Universitas Gadjah Mada, 2017.
[8] B. K. Francis, and S. S. Babu, "Predicting Academic Performance of Students Using A Hybrid Data Mining Approach," *Journal of Medical Systems*, vol. 43, no. 6, pp. 1-15, 2019, Art. no. 162.
[9] S. Ramírez, I. Triguero, and F. Herrera, "k-Nearest Neighbor -Is: An_Iterative_Spark-Based Design of the K-Nearest Neighbors Classifier for Big Data," *Knowledge-Based Systems*, vol. 117, pp. 3-15, 2017, doi: 10.1016/J.Knosys.2016.06.012.
[10] T. Winarti, D. Kerami, L. Etp, and S. A. Sudiro,"Improving Stemming Algorithm Using Morphological Rules," *International Journal On Advanced Science, Engineering And Information Technology*, vol. 7, no. 5, pp. 1758-1764, 2017, doi: 10.18517/Ijaseit.7.5.1705.
[11] T. Winarti, J. Kerami, and S. A. Sudiro, "Determining Term On Text Document Clustering Using Algorithm Of Enhanced Confix Stripping Stemming," *International Journal Of Computer Applications*, vol. 157, no. 9, pp. 8-13, 2017, doi: 10.5120/Ijca2017912761.
[12] Yang, C. Choong, C. S. Soh, and V. V. Yap, "A Systematic Approach in Appliance Disaggregation Using K-Nearest Neighbours And Naive Bayes Classifiers For Energy Efficiency," *Energy Efficiency*, vol. 11, no. 11, pp. 239-259, 2018, doi: 10.1007/S12053-017-9561-0.
[13] H. Fitriana, A. Y. Nugroho Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, C. B. Harahap "Implementation Of Naïve Bayes Classification Method For Predicting Purchase," *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 2018, pp. 1-5, doi: 10.1109/Citsm.2018.8674324.
[14] V. Krishnaiah, G. Narsimha, N. S. Chandra, "Heart Disease Prediction System Using Data Mining Technique By Fuzzy K-NEAREST NEIGHBOR Approach," Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI), vol 337, 2015, pp. 371-384, doi: 10.1007/978-3-319-13728-5_42.
[15] W. Titin, and V. Vydia, "Feature Selection for Optimizing the Naive Bayes Algorithm," In book: Engineering, Information and Agricultural Technology in the Global Digital Revolution, pp. 47-51, 2020, doi: 10.1201/9780429322235-10.
[16] S. L. Ting, W. H. Ip, and A. H. C. Tsang, "Is Naive Bayes a Good Classifier for Document Classification," *International Journal Of Software Engineering And Its Applications*, vol. 5, no. 3, pp.37-46, 2011.
[17] V. Narayanan, I. Arora, and A. Bhatia, "Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model," *Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning*, Berlin, Heidelberg, vol. 8206, 2013, pp. 194-201, doi: 10.1007/978-3-642-41278-3_24
[18] S. Hassan, M. Rafi And M. S. Shaikh, "Comparing Svm And Naïve Bayes Classifiers For Text Categorization With Wikitology As Knowledge Enrichment," *2011 IEEE 14th International Multitopic Conference*, Karachi, Pakistan, 2011, pp. 31-34, Doi: 10.1109/Inmic.2011.6151495.
[19] S. Xu, "Bayesian Naïve Bayes Classifiers To Text Classification," *Journal Of Information Science*, vol. 44, no. 1, pp. 48-59, 2018, doi: 10.1177/0165551516677946
[20] S. Xu, Y. Li, and Z. Wang, "Bayesian Multinomial Naïve Bayes Classifier To Text Classification," *Conference: International Conference on Multimedia and Ubiquitous Engineering International Conference on Future Information Technology- FutureTech 2017*, vol. 448, 2017, pp. 347-352, doi: 10.1007/978-981-10-5041-1_57
[21] R. Al-Khurayji, and A. Sameh, "An Effective Arabic Text Classification Approach Based On Kernel Naive Bayes Classifier," *International Journal of Artificial Intelligence & Applications*, vol. 8, no. 6, pp. 01-10, doi: 10.5121/Ijaia.2017.8601
[22] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, Rudy, "News Article Text Classification In Indonesian Language," *Procedia Computer Science*, vol. 116, pp. 137-143, 2017, doi: 10.1016/J.Procs.2017.10.039
[23] D. M. Diab, K.M. El Hindi, "Using Differential Evolution For Fine Tuning Naïve Bayesian Classifiers And Its Application For Text Classification," *Applied Soft Computing*, vol. 54, pp. 183-199, 2017, doi: 10.1016/J.Asoc.2016.12.043
[24] Y. An, S. Sun and S. Wang, "Naive Bayes Classifiers For Music Emotion Classification Based On Lyrics," *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, Wuhan, 2017, pp. 635-638, doi: 10.1109/Icis.2017.7960070.
[25] X. Feng, S. Li, C.Yuan, P. Zeng, Y. Sun, "Prediction Of Slope Stability Using Naive Bayes Classifier," *KSCE Journal of Civil Engineering*, vol. 22, no. 3, pp. 941-950, 2018, doi: 10.1007/S12205-018-1337-3.
[26] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani And R. Budiarto, "Evaluating Trust Prediction And Confusion Matrix Measures For Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847-90861, 2020, doi: 10.1109/Access.2020.2994222.